Efficient Training for Cross-lingual Speech Language Models

Yan Zhou [1,2,3], **Qingkai Fang** [1,2,3], **Yun Hong** [1,2,3], **Yang Feng**[1,2,3†]
[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS) [2] State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences, Beijing, China
zhouyan23z@ict.ac.cn, fengyang@ict.ac.cn

Abstract

Currently, large language models (LLMs) predominantly focus on the text modality. To enable more natural human-AI interaction, speech LLMs are emerging, but building effective end-to-end speech LLMs remains challenging due to limited data and the difficulty in expanding to more languages. In this paper, we introduce **C**ross-lingual **S**peech **L**anguage **M**odel (**CSLM**), an efficient training method for cross-lingual speech LLMs based on discrete speech tokens. We propose a novel alignment strategy that achieves cross-modal and cross-lingual alignment through continual pre-training. By conducting instruction fine-tuning following a speech-text interleaved chain-of-modality generation process, we enhance modal alignment at a finer granularity, thereby improving generation quality and reducing latency. CSLM aligns different modalities and languages simultaneously without the need for massive speech data, thus exhibiting good language scalability. Evaluations on cross-modal tasks, mono-lingual conversational tasks, and cross-lingual conversational tasks demonstrate CSLM's strong cross-modal alignment capabilities and general task abilities. [1]

1 Introduction

In recent years, the evolution of large language models (LLMs) like ChatGPT (OpenAI, 2022) has enabled the rapid development of sophisticated text-based chatbots. However, as applications of LLMs continue to expand, there is growing interest in exploring more natural human-AI interaction paradigms and unlocking the models' potential in other modalities. The emergence of speech LLMs addresses this demand, as speech-based interaction offers inherent convenience and conveys additional information beyond textual communication.

The construction of speech LLMs faces several challenges. The speech modality contains significantly more information than text modality, making speech modeling difficult. A more challenging aspect is the scarcity of speech data compared to text data, especially for certain languages. To address these issues, current researchers typically integrate text LLMs to build speech LLMs, leveraging the language capabilities and general knowledge from the text modality. The simplest approach is to cascade an automatic speech recognition (ASR) model, a text LLM, and a text-to-speech (TTS) model, but this approach brings in error accumulation and increased latency. Therefore, researchers are now focusing more on end-to-end speech LLMs. Some researchers train auto-regressive models using only speech data following the training process of text LLMS (Lakhotia et al., 2021; Hassid et al., 2023), but this approach suffers from the scarcity of speech data. Some researchers propose modular speech LLMs that establish mappings between existing speech encoders and text LLMs (Chu et al., 2023; Xie and Wu, 2024; Fang et al., 2024; Wang et al., 2024), but these methods fall short in speech generation and lack intrinsic speech-text alignment, limiting their general applicability. Other researchers explore unified modeling of speech and text (Zhang et al., 2023; Zhan et al., 2024; Nguyen et al., 2025; Défossez et al., 2024; Zeng et al., 2024) based on speech discretization techniques.

While exploring methods for modeling the speech modality, research on extending speech LLMs to more languages has also gained increasing attention. Many languages already face the problem of resource scarcity in the text modality, and this problem is even more severe in the speech modality. Building unified multilingual and multimodal representations typically requires massive amounts of data, thus how to achieve effective cross-lingual and cross-modal alignment simulta-

†Corresponding author: Yang Feng.
[1]https://github.com/ictnlp/CSLM

Findings of the Association for Computational Linguistics: ACL 2026, pages 13155–13167

neously with limited data has become a core challenge.

Recognizing the critical challenges of data efficiency and cross-lingual capability in developing speech LLMs, we propose an efficient training method for cross-lingual speech LLMs. Based on a speech LLM architecture utilizing discrete speech tokens, we introduce a novel alignment strategy that achieves cross-lingual and cross-modal alignment, and conduct continual pre-training with limited data. Subsequently, we conduct instruction fine-tuning following a speech-text interleaved chain-of-modality generation process to leverage cross-modal alignment at a finer granularity, thereby improving generation quality and reducing latency. The trained general **C**ross-lingual **S**peech **L**anguage **M**odel **(CSLM)** can align different modalities and languages simultaneously without the need for massive speech data, thus exhibiting good language scalability in terms of data requirements and training difficulty. Evaluations on cross-modal tasks, mono-lingual and cross-lingual conversational tasks demonstrate CSLM's strong cross-modal alignment capabilities and general task abilities, validating the effectiveness of the proposed method.

Unlike existing models like SPIRIT LM (Nguyen et al., 2025), Moshi (Défossez et al., 2024) and GLM-4-Voice (Zeng et al., 2024) which often require extensive data, CSLM introduces an efficient method for robust cross-modal and cross-lingual alignment. Furthermore, our novel interleaved chain-of-modality fine-tuning significantly enhances generation quality and reduces latency. Our contributions are as follows:

- We propose an efficient training method for cross-lingual speech LLMs that simultaneously achieves cross-lingual and cross-modal alignment without the need for huge amount of speech data.
- We introduce a speech-text interleaved chain-of-modality generation method for instruction fine-tuning to enhance modal alignment at a finer granularity, thereby improving generation quality and reducing latency.
- Our training method is easily scalable to other languages in terms of data volume and training difficulty, providing valuable guidance for training multilingual speech LLMs.

2 Related Works

2.1 Speech Tokenization

Speech tokenization is the process of obtaining discrete speech tokens from continuous speech waveforms. After the discrete speech tokens are extracted, they can be used like text tokens and allow for joint modeling of speech and text.

Current speech tokenization technologies primarily employ k-means, VQ (Vector Quantization), or RVQ (Residual Vector Quantization) to obtain discrete speech tokens. Hsu et al. (2021) extracts semantic tokens using k-means method on self-supervised learning representations. Chung et al. (2021) and Huang et al. (2023) obtain semantic tokens via VQ. Zeghidour et al. (2022) and Défossez et al. (2022) utilize RVQ to obtain acoustic tokens, while recently SpeechTokenizer (Zhang et al., 2024) and Moshi (Défossez et al., 2024) further use different RVQ layers to obtain both semantic and acoustic tokens. Cosyvoice (Du et al., 2024) introduces advanced speech tokenization techniques, representing speech with supervised semantic tokens derived from a speech recognition model via vector quantization, which enables semantic decoding and high-quality speech synthesis.

2.2 Speech LLM

Speech LLMs refer to LLMs that can interact with humans in speech. Depending on the modalities supported and the different approaches to modeling speech, several distinct paradigms of speech LLM have emerged, including speech-only models, modular models combining a speech encoder and an LLM, and speech-text models which jointly model discrete speech tokens and text tokens.

GSLM (Lakhotia et al., 2021) first proposes an LLM trained solely on speech, utilizing discrete speech units to train a decoder model by predicting the next token. Similarly, TWIST (Hassid et al., 2023) adopts a warm-start strategy, continuing to train a text LLM on speech data. Although these speech-only large models have the ability to model contextual relationships in speech, the sheer amount of data in the text modality naturally surpasses that in the speech modality. As a result, speech-only models can hardly achieve the same level of general task performance as text LLMs.

Qwen-Audio (Chu et al., 2023) connects a pre-trained speech encoder with a pre-trained text LLM, aligning speech representations with the text LLM to achieve speech understanding. However, this

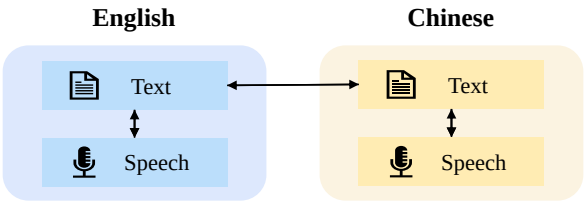

Figure 1: Alignment strategy of CSLM.

paradigm is unable to accomplish speech generation. Building on this foundation, Mini-Omni (Xie and Wu, 2024), LLaMA-Omni (Fang et al., 2024) and Freeze-Omni (Wang et al., 2024) further add a speech synthesis model after the text LLM to generate speech. These speech LLMs, which consist of a speech encoder combined with a text LLM (some coupled with a speech synthesis model), exhibit disadvantages in terms of the quality and diversity of generated speech.

Other works have attempted to jointly model discrete speech tokens with text. SpeechGPT (Zhang et al., 2023) first proposes such a method, expanding discrete speech tokens into LLM's vocabulary. AnyGPT (Zhan et al., 2024) follows this approach and improves upon the discrete speech tokens by modeling them with separate semantic and acoustic information. Moshi (Défossez et al., 2024) proposes a full-duplex working mode of speech LLM under this paradigm. GLM-4-Voice (Zeng et al., 2024) first proposes a bilingual (Chinese-English) speech LLM with massive amount of data under this paradigm, and it is capable of interacting with humans through interleaved outputs of speech and text. These models require a large amount of training data and have achieved commendable results on mono-lingual speech tasks, but their cross-lingual alignment abilities are still limited. Therefore, it is difficult for them to perform some cross-lingual speech tasks.

3 Model: CSLM

CSLM is a speech LLM based on discrete speech tokens, designed to achieve both cross-modal and cross-lingual alignment. We will introduce the architecture of CSLM, the training procedure, and the inference time workflow of this model. In addition, we will elaborate on the CSLM's possibility to be extended to more languages.

3.1 Model Architecture

CSLM consists of a speech tokenizer, an LLM, and a speech decoder. The speech tokenizer first extracts a speech waveform into discrete speech tokens, which are then modeled by the LLM to generate new speech tokens. Finally, these tokens are synthesized into a new waveform by the speech decoder.

- **Speech Tokenizer** We use the speech tokenizer of CosyVoice-300M-25hz (Du et al., 2024), which has a speech vocabulary of 4096 tokens, with a frequency of 25Hz. This speech tokenizer includes a Conformer (Gulati et al., 2020) encoder and a vector quantization module, which can transform the input speech Mel-spectrogram into discrete vectors. For training and generation efficiency, consecutive repeated speech tokens are merged before these tokens are fed into the LLM. Note that this merging operation does not introduce semantic loss, as it primarily operates on the acoustic domain.
- **Speech-Text Joint LLM** Following Zhang et al. (2023), we merge the vocabulary from the speech tokenizer with the vocabulary of a text LLM. This integration enables joint modeling of text and speech within the LLM.
- **Speech Decoder** The speech decoder consists of a conditional flow matching model and a HiFi-GAN (Kong et al., 2020) vocoder from the CosyVoice decoder, with an additional convolutional module called the duration predictor. For a reduced sequence of speech tokens, the duration predictor predicts whether each token should be repeated a specified number of times and outputs the expanded speech token sequence. This sequence is then input into the flow matching model to generate the Mel-spectrogram, which is later utilized by the HiFi-GAN vocoder to synthesize the final speech waveform.

3.2 Alignment Strategy

The goal of CSLM is to simultaneously achieve cross-modal alignment and cross-lingual alignment. The alignment strategy we designed is illustrated in the Figure 1. Within a single language, cross-modal alignment is performed between speech and text, while across different languages, alignment is achieved through the text modality.

3.3 Training Procedure

We adopt a two-stage training paradigm, which includes the continual pre-training stage and the

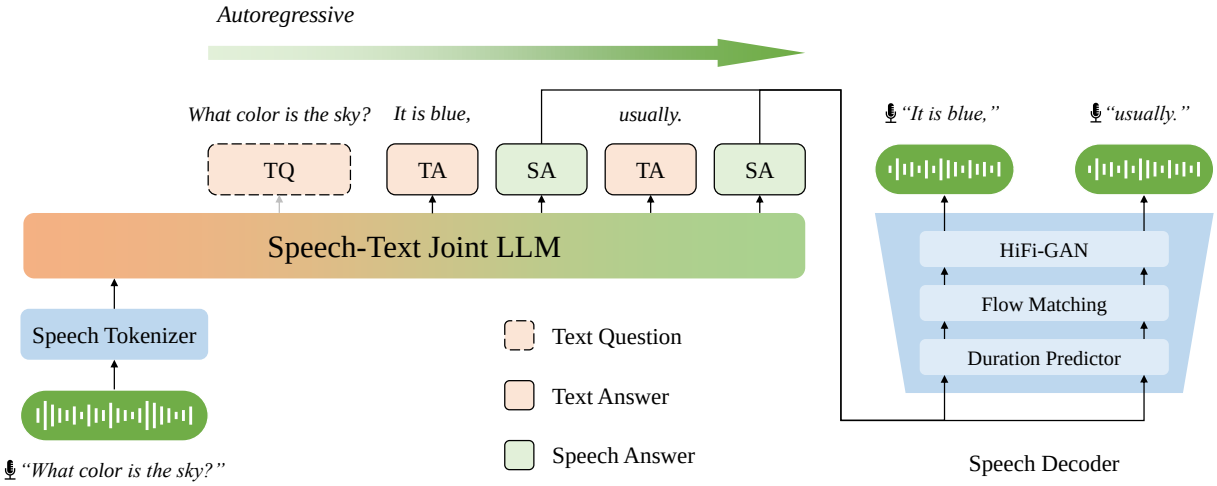

Figure 2: Model architecture and inference process of CSLM.

supervised fine-tuning stage.

3.3.1 Continual Pre-training

At this stage, we begin with an LLM already fine-tuned on instruction data, and merge the speech vocabulary with its own vocabulary. We collect parallel speech-text data in different languages to achieve speech-text cross-modal alignment. Some of the data take speech as input and text as output, corresponding to the ASR task, while the rest of the data take text as input and speech as output, corresponding to the TTS task. We collect machine translation (MT) data between Chinese and English to facilitate cross-lingual alignment. Additionally, we also collect mono-lingual instruction data in both Chinese and English, so as to reduce performance degradation in the text modality. We train the model on the aforementioned data to obtain the CSLM-base model.

3.3.2 Supervised Fine-tuning

In the second stage, the instruction fine-tuning stage, we train the CSLM-base model on text instruction and speech-to-speech conversational data, resulting in the CSLM-SFT model.

To further align speech and text and achieve higher generation efficiency, we propose a speech-text interleaved chain-of-modality based on the chain-of-modality from Zhang et al. (2023). In the original chain-of-modality, after the model receives a speech question, it generates a response in the order of text question, text answer, and speech answer, i.e., TQ $\rightarrow$ full TA $\rightarrow$ full SA. In our improved speech-text interleaved chain-of-modality, the model generates only a shorter chunk of the text answer, then immediately generates the corresponding speech answer, and this cycle repeats until the end, i.e., TQ $\rightarrow$ TA $\rightarrow$ SA $\rightarrow$ TA $\rightarrow$ SA $\cdots$.

To construct such interleaved speech-text data, we first synthesize the instructions and responses from existing textual instruction datasets into speech, and then use a Connectionist Temporal Classification (CTC) (Graves et al., 2006) aligner module to obtain interleaved responses. Specifically, for a given speech-text pair (X, Y), we first use the speech encoder of an ASR model to obtain the representation of the speech. Let $X = (x_1, ..., x_T)$ denote raw speech inputs, through speech encoder f_θ we obtain:

$$H = f_\theta(X) = (\mathbf{h}_1, ..., \mathbf{h}_T) \in \mathbb{R}^{T \times d} \quad (1)$$

where d is the dimension of f_θ. Text labels are tokenized as $Y = (y_1, ..., y_L)$ with $L \ll T$. We apply CTC dynamic programming algorithm to establish the optimal alignment path. Defining expanded label set $\mathcal{Y}' = \mathcal{Y} \cup \{\epsilon\}$ where ϵ is blank symbol, the optimal alignment path π^* is obtained via:

$$\pi^* = \arg\max_{\pi \in \mathcal{Y}'^T} \prod_{t=1}^{T} P(\pi_t | \mathbf{h}_t) \quad \text{s.t.} \quad \mathcal{B}(\pi^*) = Y \quad (2)$$

where $\mathcal{B}$ is the collapsing function that removes blanks and repeats. With such a path, we can obtain token-level alignment between the speech and text. For each token $y_l \in Y$, we can find its temporal boundaries in π^*:

$$\begin{aligned} \mathrm{A}(y_l) &= [t^{(l)}_{\text{start}}, t^{(l)}_{\text{end}}] \\ &= \min\{t | \pi^*_t = y_l\}, \max\{t | \pi^*_t = y_l\}. \end{aligned} \quad (3)$$

Based on this alignment, we organize the response data into a chunk-level speech-text interleaved se-

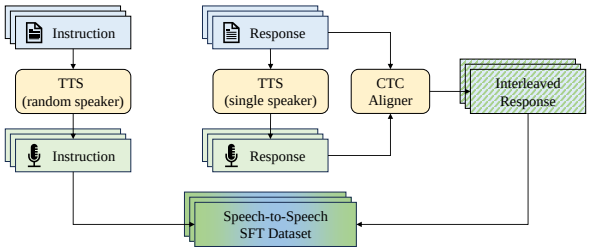

Figure 3: The construction process of the SFT dataset.

quence, with a smaller relative error than word-level interleaving. We adopt a chunk size of 7, which means when segmenting the text, we make cuts at punctuation marks unless the segment is shorter than 7 words. Appendix A shows an example of such data. It is optional whether to transcribe the speech question into text before generating the response. The total process of constructing the instruction dataset is illustrated in Figure 3.

3.4 Inference

The inference process of the CSLM model is illustrated in Figure 2. For an input speech, discrete speech tokens are extracted through the speech tokenizer, and consecutive repeated tokens are merged before being fed into the speech-text joint LLM. The LLM autoregressively outputs a chunk-level text response along with the corresponding speech tokens. The generated speech tokens are then input into the speech decoder, where they are first expanded by the duration predictor and then converted into a Mel-spectrogram by the flow matching model, and finally the audio is synthesized by the HiFi-GAN vocoder. Throughout this process, the LLM continuously outputs interleaved text and speech tokens until completion. Note that there can be a temporal overlap between playing the audio and the model generating the subsequent content [2], which significantly reduces the response latency compared to a full chain-of-modality generation.

3.5 Language Scalability

CSLM models speech using discrete speech tokens. As long as there is parallel speech-text and translation data, a new language can be integrated into CSLM's training. This enables the creation of a speech LLM capable of supporting new languages, indicating that CSLM has excellent scalability in terms of language support.

[2]More details are shown in Appendix B

4 Experiments

4.1 Continual Pre-training

Datasets In the continual pre-training stage, we continue to train the LLM with cross-modal, cross-lingual and mono-lingual text data. All the datasets that we use are open-source ones.

- **Cross-modal Data** We collect parallel Chinese and English speech-text data to form a cross-modal aligned dataset. For English, we use the English subset of Multilingual LibriSpeech (Pratap et al., 2020) dataset and the GigaSpeech (Chen et al., 2021) dataset as ASR and TTS data, with half of the examples allocated to each task. For Chinese, we use the WenetSpeech (Zhang et al., 2022) dataset as the ASR dataset and the WenetSpeech4TTS (Ma et al., 2024) dataset as the TTS dataset.
- **Cross-lingual Data** For the cross-lingual data, we select a subset from the Chinese-English translation direction in WMT17 [3], ensuring that the data count is the same for both translation directions and that the length of each example is medium (see Appendix E).
- **Mono-lingual Instruction Data** For monolingual instruction data, we utilize the InfinityInstruct [4] dataset, which includes both single-turn and multi-turn instruction data in Chinese and English.

The data statistics of CSLM's continual pre-training stage are presented in Table 3.

Model Configuration We use Llama-3.1-8B-Instruct (Dubey et al., 2024) as the foundation LLM. We expand the vocabulary by adding 4,096 speech tokens, matching the vocabulary size of the CosyVoice model.

4.2 Supervised Fine-tuning (SFT)

Datasets We train the model on mono-lingual and cross-lingual speech-to-speech instruction datasets, coupled with mono-lingual text instruction data and cross-lingual translation data for replay.

- **Mono-lingual Speech-to-speech Data** We used the text from the InstructS2S-200K English instruction dataset from Fang et al.

[3]https://www.statmt.org/wmt17/translation-task.html

[4]https://github.com/FlagOpen/Infinity-Instruct/tree/main

Task	Dataset	Model								GT
		Whisper	CosyVoice	SpeechGPT	AnyGPT	GLM-4-Voice	Moshi	**CSLM-base**	**CSLM-SFT**	
ASR	*LibriSpeech*	**2.5**	–	18.9	8.5	2.8	5.7	6.7	9.8	–
TTS	*LibriSpeech*	–	3.4	24.6	27.1	–	4.7	**3.2**	3.8	3.0
TTS	*LibriTTS*	–	**2.9**	29.1	27.9	5.6	–	3.2	4.5	2.9
TTS	*VCTK*	–	3.7	5.7	8.5	–	–	2.8	**2.7**	3.5

Table 1: Results of English ASR and TTS tasks. The test datasets include the `test-clean` set of LibriSpeech (Panayotov et al., 2015), the `test-clean` set of LibriTTS (Zen et al., 2019), and VCTK (Yamagishi et al., 2019). The Whisper model refers to whisper-large-v3, and the CosyVoice model refers to CosyVoice-300M-SFT. The last column "GT" is an abbreviation for "ground truth", representing the error rates of the original speech-text pairs from the dataset calculated using whisper-large-v3.

Task	Dataset	Model					GT
		Whisper	CosyVoice	GLM-4-Voice	**CSLM-base**	**CSLM-SFT**	
ASR	*AISHELL-1*	9.3	–	**2.5**	8.6	9.0	–
ASR	*AISHELL-2*	**5.4**	–	–	7.6	8.6	–
ASR	*AISHELL-3*	14.8	–	–	**9.2**	9.6	–
TTS	*AISHELL-1*	–	**3.3**	–	3.8	3.7	1.9
TTS	*AISHELL-2*	–	**5.0**	–	5.3	5.2	2.8
TTS	*AISHELL-3*	–	**3.8**	–	4.9	5.3	2.6

Table 2: Results of Chinese ASR and TTS tasks. The test datasets include AISHELL-1 (Bu et al., 2017), AISHELL-2 (Du et al., 2018) and AISHELL-3 (Shi et al., 2021). The last column "GT" represents the error rates of the original speech-text pairs from the dataset calculated using the paraformer large model.

Dataset	Hours	Speech Tokens	Text Tokens
	Cross-modal Data		
MLS English	44.7K	3.7B	0.6B
GigaSpeech	10.0K	0.7B	0.2B
WenetSpeech	10.0K	1.1B	0.5B
WenetSpeech4TTS	12.8K	0.6B	0.2B
	Cross-lingual Data		
WMT17 zh-en	-	-	0.6B
	Mono-lingual Text Data		
Infinity-Instruct	-	-	1.9B
Total	-	**6.1B**	**3.9B**

Table 3: Statistics of CSLM's continuing pre-training data.

(2024), along with its Chinese translation, as monolingual instruction data. We synthesize these data into speech use CosyVoice-300M-SFT [5]. More details can be found in the Appendix C.

- **Cross-lingual Speech-to-speech Data** We use the Alpaca English instruction dataset and the Chinese translation of Alpaca from Zhu et al. (2023), ensuring that each data entry has both English and Chinese versions, allowing us to create cross-lingual instructions. We continue to use CosyVoice-300M-SFT for speech synthesis. Each data entry includes bidirectional English-Chinese instruction/response pairs. The total number of such cross-lingual speech-to-speech data is 104K.

- **Mono-lingual Text Instruction Data** We randomly select a subset of 400K entries from InfinityInstruct used in Section 4.1.

- **Cross-lingual Data** We randomly select a subset of 200K entries from the WMT17 dataset used in 4.1.

4.3 Duration Predictor

The duration predictor module in the speech decoder is a two-layer convolutional module that predicts sequence durations of the input speech.

We train the model using full fine-tuning during both stages. Training details of the LLM and the duration predictor are listed in Appendix D. Details of data preprocessing are in Appendix E.

5 Evaluation

5.1 Basic Tasks

We evaluate CSLM on two basic cross-modal tasks, ASR and TTS. For both tasks, we measure the error rates compared to the ground-truth answers,

[5] https://www.modelscope.cn/models/iic/CosyVoice-300M-SFT

Model	Speech Data (k hours)
SpeechGPT*	11+
AnyGPT	57
Moshi	7,000+
GLM-4-Voice**	10,000+
CSLM	77

Table 4: Comparison of amounts of speech data in speech-text pairs used by different models. *The data volume of SpeechGPT is calculated based on the datasets listed in its paper (Zhang et al., 2023). **The data volume of GLM-4-Voice is calculated from the number of speech tokens and its frequency reported in its paper (Zeng et al., 2024).

Model	GPT Score↑		UTMOS↑	ASR-ER↓	Off-Target↓	
	T	S			T	S
En (*InstructS2S-Eval*)						
SpeechGPT	2.19	2.98	3.9	45.0	–	–
AnyGPT	–	2.31	3.3	–	–	–
GLM-4-Voice	**4.10**	**4.02**	4.0	12.8	8.5	9.0
CSLM	3.50	3.27	**4.4**	9.0	**0.0**	**0.0**
Zh (*BELLE-eval-S2S*)						
GLM-4-Voice	**4.80**	**4.70**	–	**4.6**	**0.4**	4.8
CSLM	3.78	3.37	–	6.9	0.8	5.2

Table 5: Evaluation results of SFT models on English and Chinese speech-to-speech conversational benchmarks. The "T" and "S" under GPT-Score denote the evaluation of the generated text and the transcription of the generated speech, respectively. The "T" and "S" under Off-Target denote the off-target ratio assessed for the generated text and speech, respectively. ASR-ER denotes ASR-WER for En or ASR-CER for Zh.

specifically word error rate (WER) for English and character error rate (CER) for Chinese. We use a speech decoder coupled with a duration predictor to synthesize the generated speech tokens into waveforms, after which the waveforms are transcribed back into text using an ASR model to calculate error rates, resulting in ASR-WER or ASR-CER. For English, we use the Whisper large-v3 [6] (Radford et al., 2023) as the ASR model, while for Chinese we use paraformer large [7] (Gao et al., 2022).

We compare our CSLM-base and CSLM-SFT model with the base models of SpeechGPT (Zhang et al., 2023), AnyGPT (Zhan et al., 2024), GLM-4-Voice (Zeng et al., 2024), and Moshi (Défossez et al., 2024) for English tasks, which are all speech LLMs based on discrete speech tokens. For Chinese, we compare our models with the base model of GLM-4-Voice. We also include results of specialized ASR model whisper-large-v3 and specialized TTS model CosyVoice-300M-SFT.

Model	GPT Score↑		ASR-ER↓	Off-Target↓	
	T	S		T	S
En→Zh (*InstructS2S-Eval*)					
GLM-4-Voice	**4.29**	**4.23**	**5.5**	8.5	8.5
CSLM	3.31	2.95	17.5	**1.0**	**0.5**
Zh→En (*BELLE-eval-S2S*)					
GLM-4-Voice	1.20	1.16	42.6	97.2	96.8
CSLM	**3.53**	**3.20**	**7.4**	**1.6**	**0.8**

Table 6: Results of SFT models on En-Zh and Zh-En speech-to-speech conversational tasks.

Results in Table 1 and Table 2 show our model outperforms SpeechGPT and AnyGPT, which are fine-tuned with a similar scale of speech-text parallel data. It also achieves a performance comparable to that of Moshi and GLM-4-Voice, both using speech-text pairs dozens or even hundreds of times more than CSLM. The amount of speech data we use is about one percent of that used by GLM-4-Voice and Moshi, as shown in Table 4. CSLM also performs close to the specialized smaller models. The results indicate that the CSLM base model has strong speech-text alignment capabilities in both English and Chinese.

5.2 Speech Conversation

We evaluate CSLM-SFT on mono-lingual and cross-lingual speech-to-speech conversations with both automated metrics and human evaluations.

5.2.1 Automated Metrics

For mono-lingual English evaluation, we utilize the `helpful_base` and `vicuna` subsets of AlpacaEval (Li et al., 2023), excluding examples unsuitable for speech interaction, following Fang et al. (2024). This dataset contains 199 English speech instructions and is referred to as *InstructS2S-Eval*. For mono-lingual Chinese evaluation, we select 250 instructions suitable for speech dialogue scenarios from the BELLE (BELLEGroup, 2023) evaluation set, synthesize them into audio using CosyVoice-300M-SFT and create a Chinese speech test set referred to as *BELLE-eval-S2S*. We retain the use of these two sets for cross-lingual evaluation except

[6] https://huggingface.co/openai/whisper-large-v3

[7] https://www.modelscope.cn/models/iic/speech_paraformer-large_asr_nat-zh-cn-16k-common-vocab8404-pytorch

Model	C-MOS↑				A-MOS↑
	En→En	En→Zh	Zh→Zh	Zh→En	Overall
SpeechGPT	1.83	-	-	-	3.22
GLM-4-Voice	**4.17**	**3.50**	**4.42**	3.08	**4.11**
CSLM	3.25	3.42	4.00	**4.33**	4.00

Table 7: Human evaluation results for C-MOS and A-MOS. En→X and Zh→X directions are evaluated on InstructS2S-Eval and BELLE-eval-S2S, respectively.

for instructing the model to respond in the other language. The model's speech-to-speech capability is evaluated from the following three aspects:

- **Content Quality** We use GPT4o (OpenAI, 2024) to score the outputs of the model to evaluate its ability to follow instructions and generate responses. We follow the prompts and setups in Fang et al. (2024).

- **Speech Quality** To measure the quality of the output speech, we use the UTMOS (Saeki et al., 2022) model to calculate the Mean Opinion Score (MOS), which indicates the naturalness of the English speech. We refer to this metric as the UTMOS score.

- **Speech-Text Consistency** The consistency between speech and text output is measured by calculating error rates, specifically ASR-WER and ASR-CER.

- **Language Accuracy** As a cross-lingual model, CSLM may generate results in unintended languages. We employ the metric of off-target ratio to assess this issue. See Appendix F for the calculation of this metric.

Table 5 presents the results of the mono-lingual speech conversational tasks. CSLM exhibits the best speech naturalness and demonstrates good speech-text consistency along with an extremely low off-target ratio, indicating that CSLM has advantages in cross-modal alignment and language accuracy. The content rating of responses generated by CSLM is better than that of SpeechGPT and AnyGPT. Results of cross-lingual tasks are shown in Table 6. In cross-lingual conversations, CSLM still maintains an extremely low off-target ratio. Compared to single-language tasks, there is not much degradation in content quality, demonstrating its cross-language alignment capability.

Model	Parallel Speech-Text Pairs	Random Speech-Text Pairs
SpeechGPT	1.2%	1.6%
GLM-4-Voice	39.4%	16.9%
CSLM	**72.5%**	**56.7%**

Table 8: Comparison of speech-text representation similarity on LibriSpeech `test-clean` set.

Model	GPT Score↑		ASR-ER↓	Off-Target↓	
	T	S		T	S
En→Zh (*InstructS2S-Eval*)					
CSLM	**3.31**	**2.95**	**17.5**	**1.0**	**0.5**
– *w/o MT*	3.00	2.65	25.4	4.0	**0.5**
Zh→En (*BELLE-eval-S2S*)					
CSLM	**3.53**	**3.20**	**7.4**	1.6	0.8
– *w/o MT*	3.08	2.74	9.2	**1.2**	**0.4**

Table 9: Performances of models with and without MT data on speech-to-speech conversational tasks.

5.2.2 Human Evaluations

In addition to automated metrics, we conduct human evaluations to further validate our model's performance on speech-to-speech conversational tasks. We perform a double-blind rating comparing CSLM against baseline models (SpeechGPT and GLM-4-Voice) to assess Content Mean Opinion Scores (C-MOS) and Acoustic Mean Opinion Scores (A-MOS). As shown in Table 7, CSLM achieves competitive C-MOS across both mono-lingual and cross-lingual pairs. Crucially, the overall trend of these human judgments aligns consistently with our GPT-based and UTMOS evaluations, firmly substantiating the reliability of our automated metrics and demonstrating CSLM's effectiveness in cross-lingual scenarios.

6 Ablation Study

6.1 Cross-modal Alignment Efficacy

To assess cross-modal alignment efficacy during continual pre-training, we compute the speech-text representation similarity for CSLM, SpeechGPT, and GLM-4-Voice on the LibriSpeech `test-clean` set. We measure the average sentence-level similarity in the last hidden layer for both parallel speech-text pairs and random speech-text pairs. As shown in Table 8, CSLM achieves a superior parallel speech-text similarity of 72.5%, compared to GLM-4-Voice (39.4%) and SpeechGPT (1.2%). For speech and random text pairs, CSLM scores 56.7%. This high baseline similarity for ran-

Model	GPT Score↑		ASR-ER↓	Latency(s)↓	Speedup↑
	T	S			
En→En (*InstructS2S-Eval*)					
CSLM	**3.50**	**3.27**	9.0	466.46	×2.87
– *chunk=4*	2.82	2.42	15.2	**456.88**	×**2.93**
– *full CoM*	3.21	2.92	**8.5**	1338.68	×1
– *w/o TQ*	2.01	1.20	8.9	–	–
Zh→Zh (*BELLE-eval-S2S*)					
CSLM	**3.78**	3.37	6.9	631.76	×1.92
– *chunk=4*	2.66	2.29	23.5	**620.17**	×**1.96**
– *full CoM*	3.68	**3.42**	**6.1**	1215.54	×1
– *w/o TQ*	2.12	2.00	6.6	–	–
En→Zh (*InstructS2S-Eval*)					
CSLM	**3.31**	**2.95**	17.5	437.88	×4.62
– *chunk=4*	1.62	1.55	34.7	**435.32**	×**4.65**
– *full CoM*	3.27	2.92	11.8	2024.48	×1
– *w/o TQ*	1.25	1.23	**10.9**	–	–
Zh→En (*BELLE-eval-S2S*)					
CSLM	**3.53**	**3.20**	**7.4**	666.40	×2.30
– *chunk=4*	2.74	2.45	32.1	**601.46**	×**2.55**
– *full CoM*	3.05	2.79	10.7	1531.16	×1
– *w/o TQ*	1.22	1.19	7.8	–	–

Table 10: Impact of chain-of-modality forms on speech-to-speech conversations.

dom pairs likely stems from the shared representation space developed during continual pre-training, where dense, language- and modality-agnostic embeddings inherently align text and speech. However, the substantial gap between CSLM's parallel (72.5%) and random (56.7%) scores, combined with its superiority over baseline models, affirms the specific, fine-grained cross-modal alignment established by our training methodology.

6.2 Effect of MT Data

To measure the effect of MT data during the continual pre-training stage, we conduct an ablation experiment by training a model without MT data during the continual pre-training stage, followed by same instruction fine-tuning process with CSLM. It can be observed from Table 9 that the model trained without MT data exhibits lower content quality in cross-lingual tasks. Additionally, it performs worse in terms of ASR-ER, indicating a decline in the quality of the generated speech content.

6.3 Form of Chain-of-modality

We compare chain-of-modality generation processes containing different components to validate the effectiveness of our speech-text interleaved chain-of-modality. We train a model with a chunk size of 4 in the interleaved chain-of-modality to test whether alignment accuracy would reduce performance, as alignment errors can occur at both the beginning and end of each chunk. In addition, we train a model with full CoM (i.e., generating the complete text answer and then generating the complete speech answer) during SFT to validate the performance improvement and speedup effect of our interleaving generation approach. We also train a model that skips generating the text question. Results in Table 10 show that: (i) Model with chunk size of 4 performs poorly, indicating that a low-accuracy alignment would severely damage performance. (ii) CSLM with speech-text interleaved chain-of-modality generally outperforms the full CoM model, and can bring about an average speedup of $\times$ 2.93, which validates the efficacy of this interleaved chain-of-modality. CSLM's slight advantage over full CoM model stems from its alignment granularity, which better matches the continual pre-training stage in terms of data length, as full chain-of-modality data can be very long. (iii) CSLM with no text questions shows a significant decline in metrics across all tasks, especially in cross-lingual ones, indicating that cross-language alignment occurs in the text modality, which is in accord with our alignment strategy.

7 Conclusion

We introduce CSLM, a cross-lingual speech language model. CSLM comprises a speech tokenizer, speech-text joint LLM and speech decoder, trained with an efficient method featuring cross-modal and cross-lingual alignment. Through continual pre-training and instruction fine-tuning with speech-text interleaved chain-of-modality, CSLM achieves strong cross-modal and cross-lingual alignment, enabling mono- and cross-lingual speech conversation and expanding speech LLMs' applications.

Limitations

Due to limitations in available data resources and computational resources, CSLM has not been trained on larger-scale speech and text datasets, leaving its full potential temporarily unverified. Additionally, as a cross-lingual speech model, CSLM still requires expansion to support more languages to further broaden its application scope.

Acknowledgments

We thank all the anonymous reviewers for their valuable comments on this paper. This work is supported by the grant from the Beijing Natural Science Foundation (No. L257006).

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, Changhe Song, Jiaqi Shi, Xian Shi, Hao Wang, Wen Wang, Yuxuan Wang, Zhangyu Xiao, Zhijie Yan, Yexin Yang, Bin Zhang, Qinglin Zhang, Shiliang Zhang, Nan Zhao, and Siqi Zheng. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *Preprint*, arXiv:2407.04051.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

BELLEGroup. 2023. Belle: Be everyone's large language model engine. https://github.com/LianjiaTech/BELLE.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In *Interspeech 2021*, pages 3670–3674.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. 2018. Aishell-2: Transforming mandarin asr research into industrial scale. *Preprint*, arXiv:1808.10583.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *Preprint*, arXiv:2407.05407.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *Preprint*, arXiv:2410.00037.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *Preprint*, arXiv:2409.06666.

Zhifu Gao, ShiLiang Zhang, Ian McLoughlin, and Zhijie Yan. 2022. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Interspeech 2022*, pages 2063–2067.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA. Association for Computing Machinery.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis CONNEAU, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2023. Textually pretrained speech language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 63483–63501. Curran Associates, Inc.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised

speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Zhichao Huang, Chutong Meng, and Tom Ko. 2023. Repcodec: A speech representation codec for speech tokenization. *arXiv preprint arXiv:2309.00169*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. 2024. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *Preprint*, arXiv:2411.01156.

Linhan Ma, Dake Guo, Kun Song, Yuepeng Jiang, Shuai Wang, Liumeng Xue, Weiming Xu, Huan Zhao, Binbin Zhang, and Lei Xie. 2024. Wenetspeech4tts: A 12,800-hour mandarin tts corpus for large speech generation model benchmark. *Preprint*, arXiv:2406.05763.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoît Sagot, and Emmanuel Dupoux. 2025. Spirit-lm: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52.

OpenAI. 2022. Introducing chatgpt. `https://openai.com/index/chatgpt/`.

OpenAI. 2024. Hello gpt-4o.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pages 2757–2761.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. In *Interspeech 2022*, pages 4521–4525.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. 2021. Aishell-3: A multi-speaker mandarin tts corpus. In *Interspeech 2021*, pages 2756–2760.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2024. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *Preprint*, arXiv:2411.00774.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *Preprint*, arXiv:2408.16725.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech 2019*, pages 1526–1530.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *Preprint*, arXiv:2412.02612.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *Preprint*, arXiv:2402.12226.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech large language models. *Preprint*, arXiv:2308.16692.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages. *arXiv preprint arXiv:2308.04948*.

A Interleaved Data Example

Speech-Text Interleaved Data Example

Prompt: Please directly answer the questions in the user's speech. This is input: <sosp><1490>...<1947><eosp>.
Response: [question]: Hey, can you think of a, like, really creative way to use just one single pencil?; [answer]: Use it as a plant marker in your garden,<sosp><1555>...<4450><eosp>write the name of each plant on the pencil and stick it in the soil next to it.<sosp><2937>...<4431><eosp>

Note: Content within ellipses (...) represents speech tokens.

B Temporal Overlap

We provide a concrete example of how much "temporal overlap" occurs between playing the generated audio and producing subsequent content. The question is "How do I wrap a present neatly?", and the generated answer of CSLM is:

Example of Temporal Overlap

[question]: How do I wrap a present neatly?; [answer]: To neatly wrap a present,<sosp><68><1868><1342><2323><773><2621><2554><3489><940><16><1136><2796><717><1454><760><822><2537><351><824><2110><2113><870><2110><690><822><3274><2999><2409><3887><492><2876><1688><60><302><624><eosp> start by wrapping the paper or tissue paper around the item,<sosp><1577><1446><3898><1117><124><646><3049><4><4064><3614><1122><2392><1949><51><1343><202><266><2293><489><760><822><345><740><2307><3229><409><2162><2103><101><3684><1915><1406><1698><2583><942><1122><39><1><2735><2809><2148><760><1716><1712><477><701><1618><3740><1507><39><1915><3014><1353><489><700><758><760><2515><3211><1784><870><822><3274><1404><193><800><3278><2796><1038><535><714><1404><109><33><4064><3582><3347><2230><162><eosp>...

Once the text question, the first text response sequence and the first speech response sequence, i.e., the bolded parts, are generated, the already-produced speech tokens can be used to synthesize the speech waveform, and the corresponding audio is then played. Meanwhile, CSLM continues generating the non-bolded portion, resulting in the temporal overlap.

C Mono-lingual Instruction Data

For English, we adopt the text data of the InstructS2S-200K from Fang et al. (2024). This dataset encompasses approximately 200K instruction data entries sourced from the Alpaca (Taori et al., 2023) and UltraChat (Ding et al., 2023) datasets, with instructions rewritten by an LLM and responses generated by an LLM as well. For Chinese, we utilize the Qwen2.5-72B-Instruct (Qwen et al., 2025) model to translate the InstructS2S-200K into Chinese, thereby creating a Chinese instruction dataset. Finally, we use CosyVoice-300M-SFT to synthesize speech instructions and responses. For the instructions, we use random timbres generated by the fish-speech 1.5 [8] (Liao et al., 2024) model, while for the responses we employ a fixed timbre to ensure consistency.

D Training Details

At the continual pre-training stage, we train the model with a batch size of 288 for 1 epoch. We use a cosine learning rate scheduler, where the maximum learning rate is set to 6e-5 with the first 3% of the training steps for warm-up. The maximum

[8] https://huggingface.co/fishaudio/fish-speech-1.5

sequence length of the model is 2,048. At the supervised fine-tuning stage, we train the model with a batch size of 48 for 1 epoch, and we set the maximum sequence length of the model to 4,096 and the maximum learning rate to 1e-5. The other training setups remain the same as in the first stage. All of training tasks above are conducted using DeepSpeed [9] ZeRO Stage 1 on 24 NVIDIA H800 80G GPUs.

When training the duration predictor module, we use the English speech dataset LJSpeech-1.1 [10] and the Chinese speech dataset Chinese Standard Mandarin Speech Corpus [11] from Baker, which contain 13,100 and 10,000 data entries respectively. We train the module on these datasets for 15 epochs.

E Data Preprocessing

For machine translation data in the continual pre-training stage, we filter out data where the sum of the source and target lengths is less than 128 or greater than 2048, ensuring that each example's length is medium.

We use the CosyVoice-300M-25hz [12] model as the speech tokenizer, which extracts discrete speech tokens from the waveform at a frequency of 25Hz. For the extracted speech tokens, we merge the consecutive repeated ones to improve training efficiency.

In the continual pre-training stage, each example is formatted as an instruction. Following Zhang et al. (2023), we employ GPT-4o (OpenAI, 2024) to generate ASR, TTS, and MT instructions, with a total of 10 instructions for each task. Some of these instructions are as follows.

ASR & TTS & MT Instructions

ASR (en):
Convert the following audio into written English text.
Decode the English phrases from the attached audio.
...
TTS (en):
Convert this English text into speech.
Produce English audio from the given text.
...
MT (en → zh):
Convert the English text below into Chinese.
Change the English content below into Chinese.
...

When constructing speech-text interleaved data for supervised fine-tuning, we employ pre-trained speech encoder to get the alignment. For English data, we utilize the wav2vec 2.0 [13] (Baevski et al., 2020) model, while for Chinese data, we use the SenseVoice Small (An et al., 2024) model as the CTC aligner.

F Calculation of Off-target Ratio

The specific process to get off-target ratio involves employing an external language detection tool to identify the languages present in the model's generated responses and calculating the ratio of samples that do not match the intended language. We utilize various external tools to detect the language of text responses and speech responses. For the text part of the response, we use *langid* [14] to detect the language. For the speech part, we use the SenseVoiceSmall (An et al., 2024) model.

[9] https://github.com/deepspeedai/DeepSpeed
[10] https://keithito.com/LJ-Speech-Dataset/
[11] https://www.data-baker.com/open_source.html
[12] https://www.modelscope.cn/models/iic/CosyVoice-300M-25Hz
[13] Here we refer to WAV2VEC2_ASR_BASE_960H (https://docs.pytorch.org/audio/stable/generated/torchaudio.pipelines.WAV2VEC2_ASR_BASE_960H.html).
[14] https://github.com/saffsd/langid.py